EVALUATING FLIGHTCREW PERFORMANCE: POLICY, PRESSURES, PITFALLS, AND PROMISE

Robert L. Helmreich
University of Texas

J. Richard Hackman
Yale University

H. Clayton Foushee
NASA, Ames Research Center

Draft 3.1

23 March, 1985

DRAFT

Support for the preparation of this report was provided by NASA Grant NAG2-137 and Cooperative Agreement NCC2-286 from the Ames Research Center, Robert L. Helmreich, Principal Investigator.

Contents

I. Introduction

Commercial aviation is demonstrably the safest form of mass transportation in terms of deaths and injuries per passenger miles travelled. However, despite the statistically low probability of an accident occuring on any given flight, the record indicates clearly that there is a significant incidence of sub-standard performance by flightcrews. Data to support this include reports to the Federal Aviation Administration (FAA) of accidents, near misses and other performance-related events and safety-related incidents reported under the FAA sponsored, NASA run Aviation Safety Reporting System (ASRS), which grants reporting parties immunity in most cases from Federal action for reporting errors in the aviation system. Analyses of data accumulated this system and by National Trasportation Safety Board (NTSB) investigations of accidents suggest that approximately two-thirds of accidents and incidents can be attributed to "pilot error".

Maintaining the highest possible level of flight safety is a goal endorsed by airline management, by pilots and their organizations, and mandated by the Federal Aviation Administration (FAA). Indeed, for airlines operating under Part 121 of the Federal Aviation Regulations, performance evaluation of all pilots during line operations and in the performance of standard maneuvers and emergency procedures in a flight simulator is required annually.

The environment in which commercial flight operations occur has changed dramatically since the deregulation of airline

operations in 1978. Several carriers have failed and ceased operations and new, low cost carriers have proliferated. This has led to unsettled working conditions for many pilots, pressures for increased flying time and productivity, and the hiring of substantial numbers of pilots new to air transport operations. Given the demonstrated variability in pilot performance and the extensive changes in the industry since deregulation, this would seem to be a particularly good time to take a fresh look at assessment policy and practice in commercial aviation in the U.S.

II. Context

To evaluate the strengths and weaknesses of the current system of pilot performance evaluation, we need to understand the historical roots of assessment practices and the alternatives available, including what happens in other countries. This section provides a brief look at history and selected contrasts of U.S. with foreign practices.

Historical Background

The most significant event in the history of pilot selection was World War II which created a demand for the acquisition and training of large numbers of competent airmen for both combat and support roles. The Army Air Corps created a structure for the screening, training, and evaluation of large number of candidates and accomplished this with remarkable success. Central to this endeavor was the mobilization of most of the leading American psychologists who abandoned academia to solve the practical problems surrounding the selection and training of hordes of applicants with little or no background in aviation. Much of the

research surrounding pilot selection was compiled and edited by Arthur W. Melton after the war (Melton, 1947). The volume highlights the remarkable accomplishments in program development as well as revealing the statistical limitations of research in the pre-computer era.

In all of the WW II research on pilot selection, the criterion of success was completion of or elimination from pilot training. Investigators were plagued by the fact that the criteria for elimination were largely subjective. Although attempts were made to standardize grading and to obtain ratings from multiple instructors, subjectivity in evaluator judgment was not eliminated. Forty years later, subjectivity remains a disconcerting issue for both pilots and their evaluators. While criteria for evaluating standard evolutions have improved and computers allow the precise measurement of control manipulation in both aircraft and simulators, the critical areas of judgment and decision-making are still rated subjectively with limited efforts to train evaluators in the assessment of these issues, to standardize them, and to refine a technology of evaluation.

In one of the major studies of training success conducted in 1942, the relative importance of four major categories of performance was tabulated by computing the percentages of eliminees who were cited as deficient in each. Candidates could, of course, be judged unsatisfactory on more than one dimension. The results showed the following percentages of unsatisfactory ratings for the failing group: coordination and technique - 81 percent; alertness and observation - 70 percent; intelligence and

judgment - 68 percent; and personality and temperament - 43 percent. The overall outcome of the selection research was to follow these weightings in concentration on training and to place by far the greatest emphasis on the technical, "stick and rudder" aspects of evaluation, although intelligence testing was and is included in most selection.

Although implicated in more than 40 percent of the WW II training failures, personality factors have received relatively little attention in selection research. When personality assessment is employed, its use has been primarily to screen out individuals on the basis of actual or potential psychopathology. Few efforts have been devoted to selecting in individuals on the basis of personality attributes associated with particularly effective performance.

Another theme found in W W II research is concentration on individual performance rather than crew effectiveness. The assumption was that individual skills could be combined where necessary to form effective crews for bomber and transport aircraft. Military practice was to assign pilots to categories of aircraft on the basis of judged proficiency in initial training. Those judged to be most able were channeled into single pilot, fighter aircraft while their less technically proficient colleagues were relegated to multi-pilot bombers and transports. Given the coordination and agility required for single combat in the World War 1, and the white scarf tradition of the Red Baron and Captain Eddie Rickenbacker, this philosophy was probably justified.

Current Practices in Selection and Evaluation

Selection. Not surprisingly, most U. S. airlines have built their selection practices on the military model and have used the military as the primary source of new-hire pilots. Although there is considerable variation between carriers, the norm has been for carriers to recruit pilots with a considerable amount of flying experience. At some points in time, several airlines have experimented with training pilots *ab initio*, usually selecting recent college graduates. However, no major U. S. carrier is currently using this approach.

Screening for pilot applicants typically concentrates on individual, technical aptitude, with some emphasis on personality adequacy, assessed either through psychometric instruments or interview procedures. Although candidates are being selected for a position that requires high levels of team coordination, consistent with the military model, primary emphasis is on the technical capabilities of the individual. Also consistent with the military model, when formal validation of selection procedures occurs, the criterion is performance in training.

Evaluation. As currently required by the Federal Aviation Regulations, a pilot must annually fly a series of required maneuvers, a Proficiency Check, and must undergo Proficiency Training in a simulator of the appropriate aircraft type. These assess both technical skills and individual mastery of emergency procedures. Captains must also successfully pass a "line check" which consists of observation of performance on a regularly scheduled flight. The frequency requirements for other crewmembers are less stringent. varies as a function of position

and the assessment may legally be conducted either by an FAA inspector or by a Check Airman, a pilot designated by the air carrier and approved as an evaluator by the FAA. The two possible outcomes of performance checks are "pass" or "fail" and the most severe penalty, assuming failure to pass a re-examination after additional training, is the loss of license and, hence, the right to function as a crewmember in commercial operations.

Anecdotal reports from FAA officials, Check Airmen, and other airline officials, as well as the personal observations of the authors, support a view that the Pass-Fail criteria currently employed mask a wide range of performance variability and eliminate only those who are absolutely unsatisfactory (and even these have a high probability of passing on re-examination). Given this dichotomous evaluation, the critical questions are how much acceptable variation in performance occurs among those who "pass" their requisite checks and whether these checks in fact measure the relevant dimensions of performance.

An equally important issue is that the major component of formal evaluation is the individual technical proficiency of the pilot. The formal, objective evaluation of judgment/decision-making skills and group and interpersonal skills is not currently mandated. Lacking is assessment of ability to evaluate alternatives and to make optimizing decisions in a complex, stressful environment. Also lacking is the evaluation of Captains' ability to manage the resources available and to make effective use of the human and technical support available in non-standard situations. A depressing array of accident analyses

implicate poor leadership and management of the crew as causal factors in accidents. For example, these include Captains who fail to respond to input from crewmembers indicating that their actions are seriously endangering a flight (as in the case of a Captain who disregarded repeated warnings that the fuel state was dangerously low while attempting to deal with a warning light and allowed the aircraft to run completely out of fuel and crash [NTSB, 1979]).

A recent development in training and evaluation is the limited approval by the FAA of a set of procedures called Line Oriented Flight Training or LOFT. In LOFT, a complete two or three person crew undergoes the simulation of a line flight between cities. The goal of the simulation is to reproduce the complete flight environment including dispatch releases, weight and balance computations, en-route weather, and communications with the cabin crew and with Air Traffic Control and company operations. Typically, one or more abnormal or emergency situations are introduced during the conduct of the flight.

Under a special waiver from the FAA, LOFT "training" can be substituted for part of the required annual evaluation, but assessment of performance in LOFT is mandated. This requirement creates a special set of problems in drawing a distinction between training and evaluation and will be discussed further in a later section

Overall, there is no significant public pressure for increased performance monitoring because of the statistically outstanding safety record of commercial aviation. Inside the industry, however, there is considerable awareness of the number

of "accidents that didn't happen" and of the operating costs of poor performance.

Selection and Evaluation outside the U. S.

Because the U.S. is a dominant force in international civil aviation, it is easy to assume that it is in the forefront in selection and evaluation practices. The intent of this section is to give several examples of more sophisticated practices abroad.

Lufthansa, the state airline of the German Federal Republic, is committed to the selection and training of pilots ab initio. College graduates with no flying experience are processed through a rigorous screening procedure followed, for those selected, by training from initial ground school through pilot qualification in jet transports. As Lufthansa is government owned, selection is conducted by the Federal aviation and space research institute, the Deutsche Forschungs- und Versuchanstalt fur Luft- und Raumfahrt (DFVLR). The selection process is distinctive in that it concentrates not only on an array of technical aptitudes, but also on decision-making, interpersonal skills, motivation, and personality (Goeters, 1980). Fourteen psychological factors have been isolated and are measured (Gerathewohl, 1977). A variety of assessment methods are employed including computer-controlled tasks, paper and pencil tests, and behavioral observations. As an example of the latter, the behavior of candidates playing sports is observed for indicators of interpersonal facility.

A multi-dimensional battery of paper and pencil scales is used to measure personality and achievement motivation. The scales comprising the latter were factor analytically derived and

were validated against criteria of flying performance (Kirsch, Goeters & Ewe, 1975). Multiple correlations on the order of .60 with instructor ratings of flying were obtained, along with a success-in-training rate of 92% (Witt, 1970; Goeters, 1980). German pilot selection represents a broader, more integrated, and better validated approach than is typical in U.S. civil air transport. However, even with this sophisticated approach, the selection measures were not validated against operational performance, at least in part because of labor and organizational pressures against evaluation.

With regard to the evaluation of line flight operations, British civil aviation provides a surprising example of labor-management cooperation in the service of the superordinate goals of increased flight safety and more efficient operations. The digital Flight Data Recorder (FDR) provides a sensitive, longitudinal record of critical control inputs, instrument readings, airspeed, altitude, and attitude. FDRs required on civil transports in the U.K. maintain more than forty channels of continuous information.

Beginning in the late 1970s, a working agreement was forged between the British Airline Pilots Association, the pilots' union organization, and the management of British Airways, the state-owned airline, to permit computer analysis of flight recorder data from <u>every segment flown in normal operations</u> (Mearns, 1983). The computer program employed in these analyses is designed to detect and examine "events", i.e. departures from ideal flight path and operational parameters. The current program looks for ninety-three discrete special situations or events

involving all phases of flight as well as fuel consumption. The program is also capable of recreating the cockpit instrument configuration on a video monitor in real time for detailed analysis.

Under the labor-management agreement reached, the union is charged with counseling pilots identified by the program as flying "special event"; while preserving the confidentiality of the offender. However, provisions are made for identifying to managment those involved in particularly serious or repeated events.

In addition to the safety benefits gained from the analysis of every flight, summary data are compiled on a monthly basis by aircraft type, providing management with a detailed picture of the operational performance of each fleet. These data permit timely identification of problems and implementation of changes in procedures and training. It should be noted, however, that the analyses available under the program do not allow for evaluation of cockpit management and crew coordination. In short, the British, have made a significant step in gaining more knowledge about flight performance, but have not yet achieved the goal of complete performance evaluation.

The status of flight recorder data in the U.S. is quite different. While digital flight recorders are required on widebody aircraft, those allowed by the FAA on older, narrow body trasports are primitive in comparison, recording only a limited amount of data on a few channels and allowing analog recording of data on metal foil rather than storage of digital information.

A proposed rule change by the FAA would require retrofitting digital recorders on older jet transports, particularly the Boeing 727, the Boeing 737, and the McDonnell-Douglas DC-9 (FAA, 1985). The change, if initiated, would still leave the U. S. behind Britain in the extent of data captured and in the fact that data are only utilized in the event of accidents or reportable incidents.

Similarly, Cockpit Voice Recorders in the U.S. record a single channel of information on a continuous loop, thirty minute tape that inputs from a single area microphone. In the case of voice data, recording quality is frequently poor and it is often impossible to determine who said what. The British, in contrast, record data from each flight position on separate channels, the result being good quality records allowing easy identification of the source of each communication. Whatever the quality of the flight and voice records in the the U.S., the data are only examined after accidents or incidents. Both air carriers and pilots and their organizations have opposed improvements in flight data recording. Airlines have resisted the new technology on grounds of cost and fears of greater vulnerability to disciplinary action by the FAA. Pilots have used a variety of arguments to oppose greater access to flight data, with a particular goal being the protection of members from disciplinary action by either the government or the carrier. In many ways, reactions to the FDR illustrate many of the pressures surrounding the broader issue of overall flightcrew performance evaluation that will be explored next from the perspectives of the involved parties.

Section Implications

This section has established the status of flightcrew evaluation in the U.S. by giving a brief review of historical developments in pilot selection. As a counterpoint, examples of selection and evaluation outside the U.S. were described. While the majority of the discussion will be directed at evaluation practices, it is important to keep in mind the selection practices described above while considering evaluation policy and techniques.

III. Perspectives on Crew Assessment.

Although logic suggests that everyone associated with the air transport system would favor the most precise and rigorous evaluation of pilot and crew performance, there are costs as well as benefits associated with increasing the stringency of assessment standards. In this section we will examine assessment practices from the perspectives of management, pilots, and the Federal Aviation Administration.

Management Perspective.

The costs in terms of public confidence to an airline from a crash are obvious. Not only is there a loss of revenue, but also increases in insurance costs. In addition to benefits from safety-related concerns, flightdeck performance optimization should also result in decreased aircraft fuel and maintenance costs. Despite the demonstrable benefits from improving flightcrew performance, U.S. airlines have been notably non-aggressive in seeking more comprehensive evaluation of flight

behavior and in striving for higher levels of crew performance. There appear to be three basic causes of this behavior. The first has to do with the nature of the deregulated industry, the second with the preservation of harmonious labor relations, and the third with potentials for liability from maintaining records reflecting variance in competence. The three factors are interrelated in many ways but will be treated independently.

The impact of regulation and deregulation. Until 1978, both the routes flown by individual carriers and the fares charged were controlled by the Civil Aeronautics Board. During this period, carriers were given generally non-competitive routes and passenger fares were federally controlled to provide a "reasonable rate of return" to the airline, even including subsidies for carriers flying to certain destinations where traffic was light. There was little incentive to contain costs because these could be passed on to the passenger with Federal blessing, as return rather than efficiency of operations determined the fares charged for a particular route. In this operational environment, the major threats to profitability were prolonged labor disputes which could divert passengers to alternative methods of transport and the general state of the economy, which could determine the extent of air travel.

Under an initiative of the Carter administration, civil air transport in the U.S. was deregulated in 1978. The effect of this was to place airlines for the first time in a fully competitive environment where routes were freely available and where fares, and commensurately profits, would be determined by the free play of the marketplace.

One of the outcomes of deregulation may have been a decrease in the level of communication on safety-related issues among air carriers. In a highly competitive environment, the pressures to avoid exchanges which could provide a competitive advantage tend to have a chilling effect on dialogue. In addition, the financial pressures imposed by the new environment have perhaps reduced the commitment air carriers are able to make to research and development relevant to overall system safety. These concerns were recently reflected in testimony by James Burnett, Chairman of the National Transportation Safety Board (NTSB) to the U. S. House of Representatives Transportation Appropriations Sub-committee. Burnett cited reduced training and equipment expertise in "undercapitalized startup carriers" and the likelihood that these factors will negatively effect safety. No systematic studies have been conducted that would allow verification of these possibilities. In any event, however, these new pressures on the system undoubtedly place more responsibility on agencies such as NASA to provide the research data necessary to maintain and enhance the effectiveness of commercial aviation.

Performance evaluation and labor relations. U.S. pilots and their professional organizations have opposed increases in formal pilot evaluation. After deregulation became fully in place in the early 1980's, the major airlines began to feel severe financial pressures from competition with newly formed, low cost carriers and to face mounting losses from a downturn in the economy. Pressed by financial problems, many airlines began negotiations with their workforces, especially pilots (the highest paid

group), seeking significant concessions in wages and work rules. Any pressure to increase pilot performance evaluation would have been likely to upset the tenuous relations between management and pilots, As either contract intransigence or a strike could have had disastrous effects on already financially shaky carriers, anything, including evaluation, that could upset relations was far from the forefront of management concerns.

The newer, low cost airlines, on the other hand, having already obtained a pilot force willing to work longer hours and undertake more varied responsibilities for less pay, were not motivated to upset this profitable and productive state by imposing performance evaluation standards more rigorous than the established airlines. The end result has been that carriers have stayed clear of evaluation issues, complying only with Federally mandated standards.

Potential for liability. A seldom verbalized but salient consideration for management is the potential of crew performance data to increase a carrier's liability in the event of an accident. For example, if an accident were determined to be a result of "pilot error" and it were further determined that the assessed performance of involved crewmembers was below the carrier's average in previous evaluations, litigants could argue that the airline callously endangered passengers' lives by boarding them on a flight staffed by substandard personnel. An example of this is found in the case of an Air Florida jet which crashed into a bridge on takeoff in Washington, D.C. (NTSB 1983). Pilot judgment and performance were determined to have been causal factors and it was further disclosed that the Captain

involved had failed a Proficiency Check prior to the accident (although he had passed the examination after retraining). It is impossible to determine the precise impact of this disclosure on the outcomes of lawsuits and the subsequent failure of the airline, but the effects were clearly negative.

In addition to the forces just described which would clearly dictate caution in undertaking a policy with considerable potential costs, the backgrounds of top management personnel may also reduce the level of concern with performance evaluation. Unlike the early days of air transport when many chief executives were themselves pilots or advanced through the operational side of the industry, many of the leaders in today's environment come from a background in finance or marketing. Given the overall safety record, such individuals could be expected to de-emphasize the importance of performance issues in the face of other considerations.

Pilots' Perspectives

As we have noted, pilots have generally opposed changes in the nature and scope of performance evaluation and have resisted attempts to increase the technological sophistication and utilization of information from Flight Data Recorders and Cockpit Voice Recorders.

There are conflicting interests for both pilots and their representative organizations. Obviously, it is in pilots' personal and professional interests to achieve the highest degree of safety and to promote the financial health of their employers through enhancing efficiency of operations. On the other hand,

negative performance evaluations can result in loss of license and professional livelihood.

At first glance, it would appear that opposition to more comprehensive performance evaluation represents a triumph of narrow self-interest in preserving jobs over more general values. However, a number of concerns with the equity of evaluation are well founded. Dating back to the earliest research on performance and selection during World War II, subjectivity in evaluations has been a problem. The recent growth in emphasis on judging the decision-making and managerial skills of Captains as well as overall crew coordination has made this issue even more salient, given the subjective nature of these concepts. The technology of evaluation and the training of evaluators have not advanced enough to provide reassurance to those being evaluated that they are being judged by a reliable, valid, and impartial system.

Adding to the evaluation anxieties of pilots is the fact that labor-management relations between pilots and airlines have been more adversarial than collegial in recent years. There is a perception among pilots that management could use evaluation as a club to bring pilots into line. It is argued that subjective evaluations could be used to eliminate those who are particularly effective spokesmen for pilot concerns and that, for example, Cockpit Voice Recordings could be used as subtle blackmail to stifle dissent.

The Perspective of the Federal Aviation Agency

The FAA, as the responsible Federal agency, is charged with mandating practices which will ensure the highest level of safety in commercial aviation. However, the FAA necessarily responds to

a number of conflicting pressures. While safety is a paramount concern, the Agency is also cognizant of the need to promote civil air transport and is sensitive to pleas from carriers regarding the fiscal impact of regulations. It is also clearly aware of the lobbying pressures brought to bear by of pilots' organizations which argue that their constituents may be harmed by restrictive regulations. Usually on the other side of most safety-related issues are passenger groups which lobby for more rigorous controls and evaluations.

The strongest pressures for more stringent performance measurement and evaluation come from the National Transportation Safety Board, the Federal agency charged with determining the causes of accidents and recommending procedures to avoid the reccurence of similar events. Based on its objective interpretation of data from a number of air transport crashes, the NTSB has been recommending to the FAA for some years that it increase requirements for data capture in Flight Data Recorders and Cockpit Voice Recorders and that there be increased emphasis on training in assertiveness for junior crewmembers and in crew coordination for all cockpit crews.

Despite the weight of objective data, the FAA has been slow in accepting the recommendations of the Safety Board and the status of evaluation has remained essentially unchanged. Given the conflicting pressures to which it is exposed, the FAA is not likely to become more aggressive in its regulatory role in the forseeable future.

Section Implications

This section has stressed the point that there are several legitimate players in the game of assessment, each with certain special interests and blind spots, but each also with special influence and resources that can be used to work toward more satisfactory assessment policies and practices. Having established these viewpoints, we turn next to what we view as the major challenges in assessment and to some ideas about what might be done to meet them.

IV. Assessment Challenges for the Next Decade

This section focuses on the aspirations for assessment that have been identified in our review of the status of evaluation in the air transport system. These include methods and policies that do not cover over real variance in performance, and policies and techniques that include much more than stick and rudder flying - especially judgment/decision-making and group (crew) level issues. We will identify three major challenges which, when taken together and integrated, offer considerable hope for improving the scope and effectiveness of individual and crew performance evaluation.

Capturing and Using Extant Data

An enormous amount of data on performance in the air transport system already exists. The "hard", archival data primarily reflect deficiencies in performance. These include records of failed checks, accident and incident reports, and the vast database accumulated by NASA's Aviation Safety Reporting System. While these sources of data have great heuristic value in

studying problems in the system, they are of limited value in understanding the variability in crew performance under normal circumstances.

There are, however, extensive informal evaluation systems among operational personnel in the air transport system that reflect the true variability in capabilities and performance. "Real" evaluation tends to occur unsystematically and in a manner invisible to researchers and the regulatory agency. Because of limitations on the scope and range of pilot evaluations specified by the FAA, Check Airmen and other airline officials tend to maintain their own evaluation system based on their subjective, expert knowledge of pilots. For example, the Vice-president of a major airline told one of us that he kept a notebook with his evaluation of the strengths and weaknesses of the Captains in his jurisdiction. Other Check Airmen we know maintain lists (formal or informal) of the "good" and "bad" pilots they have evaluated. These judgments fall outside the system because they tend to be global and subjective and not clearly specified as part of the proficiency criteria of the formal system. Check Airmen know that Captain X is a "bad pilot", but they cannot fully articulate the reasons for the judgment. Typically, though, the sources of the evaluation lie in observations of poor decision-making, deficient crew coordination, and poor communication on the flight deck rather than lack of "stick and rudder" proficiency.

In our observations, we find that there is a considerable consensus among Check Airmen as to who is and isn't a good or bad pilot along these dimensions. It should be noted, however, that

these informal evaluation systems focus on the individual rather than on the performance of crews, reflecting along with the historical emphasis on individuals rather than teams, the transitory nature of crew pairings.

The existence of an "invisible" evaluation network outside of formal boundaries can provide information of great use to those managing an organization. However, this same informal process can hinder the development of a more precise, formal evaluation system that would encompass the broader issues related to team performance and team leadership. The judgments of overall ability made by Check Airmen are usually not based on objective indicators of performance but rather on judgments of the processes involved in flightdeck management. Because these evaluations cannot be related to objective, behavioral criteria, because Check Airmen feel that they can subjectively recognize good and bad crew performance, and because the informal system has proved somewhat useful, little or no pressure has been exerted on airlines or on the FAA to refine evaluation procedures and to modify the formal system to incorporate these issues.

The Conflicted Role of the Check Airman. As discussed above, Check Airmen are the repository of much of the data needed to improve evaluation technology. Because this role is the pivotal one in flightdeck performance evaluation, it may be useful to discuss it in more detail.

This position, as it has evolved, places the incumbent in a uniquely uncomfortable position in relation to management, peers, and the FAA. The Check Airman is, basically, a line pilot who has been judged to have the requisite skills to evaluate the

performance of fellow pilots. After evaluation of his or her qualifications, the Check Airman is certified by the FAA and acts as the Agency's surrogate in evaluating the company's pilots for initial qualifications, continued proficiency, and upgrade to more senior positions. From the perspective of flightcrew members, the Check Airman is a fellow pilot charged with the difficult task of passing judgment on peers in the interests of flight safety and high professional standards. From the management viewpoint, he or she is both a pilot and a member of management, someone whose loyalties should be both to the maintenance of flight standards and to the success of the organization. In the eyes of the FAA, the Check Airman is an individual who, despite his designation as an agent of the regulatory agency, may be biased in evaluation toward the protection of peers or the protection of the organization. Line pilots tend to regard the Check Airman as the representative of both the FAA and company management and also as the most direct threat to maintenance of their licenses and professional livelihood.

The authors have observed Check Airmen conducting evaluations at several major airlines and can offer some generalizations from these observations. One is that through selection and surveillance, Check Airmen as a class are extremely well-qualified technically and highly motivated to be fair and accurate in their assessments. They are also highly aware of their personal responsibilies for aviation safety. They are equally aware of the conflicting pressures and responsibilities

that the role engenders.

Another is that the training and standardization of Check Airmen in evaluation techniques are quite limited. While there is overwhelming agreement on what constitutes unacceptable performance in the technical, stick and rudder aspects of flying, there is no similar consensus on evaluation of areas such as decision making, crew coordination, and overall cockpit management. The latter factors, however, have been isolated as causal factors in a number of transport accidents and are major sources of concern both to regulatory bodies and to air carriers. There is increasing pressure on Check Airmen to evaluate performance in these areas. This creates a serious dilemma for the evaluator who is aware of the need for such judgments to maintain aviation safety and equally aware of the lack of valid guidelines for making such judgments objectively. In practice, most Check Airmen seem to base "unsatisfactory" ratings on technical performance and to be reluctant to downgrade pilots in the more subjective areas. Many express serious concerns about this, acknowledging that they do observe pilots who are seriously deficient in crew coordination and cockpit management but are unwilling to fail them on these grounds.

Exploiting Available Technology

As we have noted, required technology regarding the recording of flight parameters and verbal interactions in the cockpit is still relatively primative in the U.S., at least in contrast with the state of the art and required instrumentation in some other countries. Additionally, the assessments made during formal evaluations are quite limited in scope,

concentrating on the technical rather than judgmental aspects of flight management. However, one relatively recent innovation does hold great promise for improving evaluation and performance. This is the development of Line Oriented Flight Training or LOFT.

LOFT: Pitfalls and Promise. The introduction of Line Oriented Flight Training (or LOFT) has been perhaps the major development in training in recent years. Using modern simulators and carefully crafted scenarios, crews can experience all aspects of line flight operations including bad weather, communications with ground and cabin crews, and the full range of abnormal and emergency flight conditions. Aviation psychologists, especially those associated with NASA, have become heavily involved with the development of LOFT and have developed guidelines aimed at maximizing the psychological and training impact of the experience (Lauber & Foushee, 1982).

Even highly experienced crews report that LOFT is a powerful training tool that allows them to test all their skills, both technical and managerial, under extraordinarily realistic conditions. While crews can gain many valuable insights from the experience itself, especially when the simulation is videotaped and can be reviewed, what is missing is a set of meaningful criteria of both process and outcomes which would allow the instructor to provide detailed feedback and training for participants.

The situation becomes more difficult with regard to formal appraisal of crew performance. Although conceptualized as a training tool, the FAA, in approving the substitution of LOFT for

one of the required annual checks, instituted the requirement that performance must be "satisfactory", i.e. must meet the general standards applied to evaluation of individual pilots in a simulator or line check. This requirement poses great difficulties for those conducting the training. Other than the usual technical proficiency standards that can be applied to the pilot manipulating the controls, there are few valid guidelines for evaluating the performance of the crew as a whole and for partitioning the blame (or praise) among individual crewmembers. Adding to difficulty in evaluation is the fact that, given the basic problem of flying from Point A to Point B safely, while coping with particular situational issues, there is no single best way to conduct the flight safely and expeditiously. Different crews capitalizing on their own experience and knowledge of capabilities and limitations may come up with quite different, but equally effective solutions to the problems encountered. This poses a difficult problem for the evaluator which will be discussed in a later section.

In general, Check Airmen have been extremely reluctant to give "unsatisfactory" ratings for LOFT, using the argument that "if the crew found it a significant learning experience, it was a satisfactory session regardless of the performance exhibited".

In fact, the LOFT setting provides perhaps the most valuable resource available for performance evaluation. Not only can variations in response to a complex, but standardized situation be observed, but also scenarios can be constructed especially to test specific aspects of behavior. For example, an individual may be adquate in all the technical aspects of flying and may be

functioning well as a co-pilot. If there are questions, however, about the pilots capacity to fill the Captain role, a scenario may be constructed to allow him to demonstrate decision-making and managerial skills while serving as Captain on the simulated flight.

In general, LOFT is proving to be extremely useful in the training function and is gaining wide acceptance. Awareness of the utility of the LOFT paradigm for both normal and special evaluations is growing, but its potential has only begun to be realized.

In the opinion of the authors, the LOFT approach can provide air transport with the best possible approach to both training and performance evaluation, but only if several developments occur. The first is the evolution of an assessment technology that is accepted by operational personnel as being reliable, valid, and objective. The second is the achievement of a reduction in the pressures against evaluation operating on both the airline management and line pilot groups. In the following sections, we will discuss specific problems in performance evaluation, research needed, and possible modifications in evaluation procedures.

Developing Better Tools

As we have noted above, there are real opportunities for improving assessment in already existing techniques such as LOFT. However, progress will be limited if we are restricted to the devices and data that are currently available in the U.S. The following are some developmental possibilities that strike us as

promising.

Technology. The British model of not only requiring highly sophisticated Flight Data Recorders and Cockpit Voice Recorders, but utilizing at least the FDR data from routine operations for system and individual evaluation seems extremely valuable. With due regard for protecting the rights of individuals, such data could provide important, timely information on the performance of aircraft and crews.

With the evolution of new, more automated aircraft there are undoubtedly a range of possibilities for capturing more comprehensive information on the control and management of flight. To achieve this, the case must be made that it is in the interests of both individuals and organizations to achieve more complete knowledge of performance.

Assessment Methods. The assessment of crew performance is greatly complicated by the fact that there are multiple ways to achieve good outcomes. Different crews faced with an in-flight emergency situation may employ a variety of different problem solving and management techniques, all of which may effectively resolve the situation and, accordingly, must be considered good solutions. This makes it extremely difficult to define objective criteria for "good" team performance in complex situations and difficult to train evaluators to achieve high reliability in assessment.

Given the fact that a variety of behaviors may all result in satisfactory outcomes and that there is relatively little variability in the safety of flight, the burden of evaluation in most instances involving team coordination and performance must

fall on the processes involved in the conduct of flight operations. Because much of flightcrew evaluation centers on how crews manage non-standard, critical events, it may be useful to draw a distinction between acute and continuing abnormal situations. Acute situations are ones that require immediate action, usually on the part of the pilot physically flying the aircraft. Such action may range from relatively simple changes in flightplan, such as an order from Air Traffic Control to change course or altitude to emergencies such as the need to avoid a mid-air collision, to abort a take-off or landing, or to cope with other immediate threats to safety. Typically, these situations involve procedures which are heavily overlearned. The key factors are the recognition of the need for action and the process of smoothly and efficiently executing a prescribed strategy. The evaluation of performance in acute situations is relatively straightforward as it involves timing and execution of defined action. Many of the behaviors assessed in the current Performance Check fall into this category.

Continuing situations, on the other hand, are those where conditions require decision making involving consideration of alternative courses of action and the development of a strategy of action. These are situations which are not overlearned and where only general training and experience is relevant. Examples of this type of situation would include mechanical malfunctions that do not pose an instantaneous threat but place in jeopardy the safe continuation or completion of a flight. These are the types of problems that require the coordinated action of the full

crew and are those most likely to be included in the development of scenarios for LOFT. They are also, not surprisingly, the kinds of situations frequently encountered in incidents and accidents where conclusions of "pilot error" are reached. It is in this area that the methodology of assessment is most deficient.

Hackman (1983) has developed a normative model of group effectiveness which posits that the overall effectiveness of a work team is a joint function of: 1. the level of effort group members collectively expend carrying out task work; 2. the amount of knowledge and skill members bring to bear on the group task; and 3. the appropriateness to the task of the performance strategies used by the group in its work. The nature of leadership is also critical in determining team performance. Hackman has specified three leader functions that can improve team performance: 1. attempting to put into place the kinds of structures and systems (including rewards, training, information, and other resources) that will provide a working environment that supports competent member behavior and excellent team performance; 2. building the team and fostering collective learning; and 3. providing direction to manage the task-effective interaction among members. This model would seem to fit well the flightdeck environment, especially with its greater stress on the processes of goal attainment than on variations in performance outcomes. What is conspicuously lacking, however, is an empirical research base showing the fit of this model to conditions in the cockpit.

The study of small groups and group performance has lagged far behind other areas in psychology in both the development of a

solid empirical base and in the application of sophisticated methodologies and analytic techniques (Helmreich, Bakeman & Scherwitz, 1973; Helmreich, 1975; McGrath, 1983). Caught between the artificiality of the laboratory and the complexity of natural situations, investigators have directed more attention to individual assessment and less to understanding group performance, a situation that parallels the process of evaluation in aviation. An improved conceptual understanding of the processes of group task enactment is needed along with methodologies useful for the practical assessment of group effectiveness. These issues converge on the flightdeck when evaluators are faced with the need to assess both individual and crew performance.

The necessary first step in establishing more effective flightdeck performance evaluation must be research aimed at the development and validation of more precise indicators of individual and group performance. Fortunately, the aircraft simulator is a superb research environment, as demonstrated in a seminal NASA investigation. In this study, Boeing 747 crews flew a full mission scenario involving a trans-Atlantic crossing with several mechanical malfunctions necessitating complex decisions about aborting the flight and returning to the point of origin in deteriorating weather conditions (Ruffell Smith, 1979). The study was unique in capturing and analyzing not only technical performance, but also verbal interactions in an attempt to relate management styles and behaviors to objective performance (Foushee & Manos, 1981). The Ruffell Smith and related studies were

instrumental in demonstrating many of the limitations in overall crew performance assessment, and, more importantly, gave impetus to a major research endeavor by NASA in the area of crew performance. A major aircraft simulator research center is nearing completion at NASA-Ames Research Center dedicated to research on crew performance (Foushee, 1984). Not only does this setting allow applied research on determinants of flightcrew performance, it also provides an outstanding environment for the exploration of basic questions about small group behavior.

The attack on team performance assessment must be multidimensional, including observations and ratings in unconstrained, line operations and in controlled flight simulations which present the same operational problems to a number of crews. It is in this aspect of refining performance assessment that the knowledge and expertise of Check Airmen, who are most intimately aware of the variations found in the system, can be most useful. Given a commitment to research and to the protection of individuals and organizations, Check Airmen can provide invaluable data on the critical parameters in flightdeck management which can greatly facilitate the research needed to improve assessment technology.

An important element in the research approach involves the development of multiple coding schemata designed to capture the molecular aspects of performance enactment. Coding categories are evaluated using time-lined videotapes of LOFT scenarios flown by line crews. Three broad areas are specified, information transfer, control, and group climate. Information transfer components include both operational and social-emotional

communications and further include both breakdowns of the relative contributions (initiated and reactive) of team members and the qualitative aspects of the interaction (i.e. the forms of communication). Control factors consist of direct and indirect attempts to influence and "manage" the ongoing situation. Climate refers to indicators of the affective tone of group interactions and the inferred states of individual team members. No attempt has been made to impose independence on the behavioral categories; they are related cuts of the same phenomena.

Process variables such as those just described are difficult to interpret except within the context of the task situation. For this reason, several different frames of reference are being explored. The most basic consists of examining each phase of flight (pre-flight, take-off, climb, cruise, descent, approach, and landing) discretely and, within each phase classifying the situation at points of measurement as being normal, acute non-standard, or continuing non-standard. Another approach involves classifying activities in terms of their relationship to necessary actions during each phase of flight. That is, actions may be directed towards coping with the immediate situation, may be attempts to complete activites that should have been accomplished earlier but were deferred, or may be focussed on future actions and the development of action strategies. A final approach consists of utilizing leader (Captain) behavior as a benchmark against which to measure the behaviors of the other team members.

With the development and validation of reliable procedures

for assessing crew performance in line situations, several important issues can be addressed. One involves the assessment of the effects of different training techniques, especially those dealing with crew coordination and cockpit resource management, on flightcrew performance.

While the performance appraisal systems employed in research will doubtless continue to be complex and time consuming to administer and evaluate, less detailed variants should capture the critical aspects of performance and provide meaningful comparisons with norms. A particularly important application of the research should be the development of evaluation procedures that can be reliably employed by Check Airmen and others charged with performance assessment, including training personnel.

Section Implications

There is great potential for improvement in assessment/evaluation policy, tools, and practices. We feel that the protected use of available data on flight parameters along with new technologies for recording relevant information can offer many benefits. However, the greatest promise lies in the enhancement of LOFT technology. The ability for crews to learn their own strengths and limitations in a controlled situation is invaluable and similarly, the setting provides an unparalleled venue for evaluating overall crew performance as well as individual capabilities. However, the practical utility of LOFT as well as individual and group evaluation on dimensions of judgment/decision-making and interpersonal effectiveness must proceed hand in hand with refinement in the technology of assessment itself.

V. A Modest Integrative Proposal.

Several modifications in the current checking process could serve to optimize the training potential inherent in the LOFT approach and to obtain the greater precision and relevance of evaluation needed. The following are suggested as procedures which could improve both capabilities.

Redefining LOFT.

The first step should be to separate formally the training and evaluation functions of LOFT. We feel that it is invaluable for crewmembers, especially Captains, to receive the opportunity available during LOFT to gain more understanding of their behaviors and their consequences and to be able to explore new behavioral strategies in a "no-fault" situation. On the other hand, we feel that evaluation of performance within a line oriented paradigm provides information that is unavailable in the Proficiency Check and unlikely to show up in most Line Checks.

We recommend that the term LOFT be reserved for training periods where formal, mandated evaluation is explicitly omitted. It is suggested that each crewmember have the opportunity to participate in a LOFT at least once a year, flying in his or her normal position. LOFT sessions should be run by specially trained Instructors or LOFT Coordinators who can provide extensive feedback on on both individual proficiency and group process and can serve as resources for additional training in either resource management or technical areas. It is probably preferable for the LOFT Coordinators not also to be Check Airmen, as it is extremely

difficult to switch between the training and the evaluation modes. Of course, if the LOFT Coordinator observes serious problems, individual or interpersonal, that could threaten safety, there is an implicit requirement that this behavior be reported. This, however, is no different from the responsibility any Check Airman or even any line pilot should feel for the maintenance of flight safety.

LOCK: A New Approach to Checking.

We believe that the evaluation of all crewmembers in the environment of line operations presenting abnormal situations requiring team coordination is also essential to flight safety. It is recommended that each flight crewmember be observed in this setting under formal evaluation conditions. This is particularly important for the Captain, as his or hers is the primary responsibility for decision-making and crew coordination as well as for the overall management of the flight. To keep the distinction between this type of evolution and LOFT as distinct as possible, we recommend the use of a different terminology. One label that could be applied to the formal evaluation is LOCK or Line Oriented CheckK. The LOCK would be conducted by a Check Airman qualified to give simulator Proficiency Checks who has also been given formal instruction in both the conduct of line oriented simulations and techniques of individual and group evaluation.

We feel that reliable and objective assessment of overall crew performance can be achieved in the LOCK. In the case of the Captain, valid individual evaluation can also be accomplished because his or her responsibilities include the coordination and

utilization of all available resources, especially including the capabilities of other crewmembers. First Officer performance is more difficult to evaluate in the full crew context. The First Officer role is a pivotal one. He or she is expected to be technically proficient in the operation of the aircraft, in fulfilling an important support role, and also able to assume the Captain's role in the event of subtle or profound incapacitation. That the latter is an important issue is seen in research on First Officer reactions to Captain's incapacitation (Harper, Kidera, & Cullen, 1971), in accidents where the First Officer has failed to assume responsibility (e.g. NTSB 1980), and in the fact that a considerable number of FAA approved LOFT scenarios involve Captain incapacitation. It would appear to be important to evaluate First Officers' performance in the Captain role and we recommend that they be evaluated in a LOCK where they enact the Captain role, although the performance standards in this case might not be as stringent as those applied to line Captains.

If LOCK is to be used effectively and fairly, particular care must be exercised in the development and evaluation of the scenarios employed. Every scenario should be tested to verify that it elicits behaviors relevant to cockpit management and that it does not place unfair pressures on the crew being evaluated.

It is particularly recommended that LOFT and LOCK form integral parts of both initial and upgrade training for all flightcrew positions. Especially in initial training, LOFT provides a means of teaching crewmembers the characteristics of the operational environment in which they will be flying. LOCK

can provide more assurance that the individual is capable of handling the multiple demands of the line setting than any individual check.

The implementation of LOCK assessment is not intended to replace training, practice, and evaluation of the traditional, "stick and rudder" pilot skills that are measured in the present Proficiency Training and Proficiency Check. LOFT and LOCK are seen as a supplement to current practices.

Issues in implementation.

To adopt the type of crew and individual evaluation discussed here would require a change in the Federal Aviation Regulations. This necessarily would involve a redefinition of pilot proficiency and a recognition of the interdependency of crewmembers in achieving safe and effective flightdeck management. This is an endeavor the FAA is only likely to undertake if both airlines and pilots support such innovation. The implication of this is that the evaluation process must be cost effective in terms of enhancing flight safety and efficiency, must be objective and equitable, must not subject pilots to the risk of losing their licenses as a result of capricious or biased evaluation, and should not increase the liability of airlines for pilots' errors.

We feel that this approach to team evaluation is of great potential value and is much needed in commercial aviation. We also believe that it is realistically attainable, given a good faith approach to safety enhancement by all concerned parties. It would, however, be a major strategic error to press the FAA to mandate this type of evaluation for all carriers at this point in

time. This is particularly true before more progress has been made in developing and validating assessment techniques. The strategy with the highest probability of success is likely to be to allow organizations, by waiver, to institute such an evaluation program on a trial basis. If the programs achieve their potential, the force for implementation will come from below rather than above.

In the initial, "pilot" phases of utilizing this approach to flightcrew evaluation, the critical need will be for databases that can be used to validate criteria in line operations and to determine their utility for development of selection criteria and training procedures. A method should be developed to maintain the confidentiality of individual pilot records (above and beyond the current "satisfactory" - "unsatisfactory" evaluation) so that these more detailed performance measures can be used to improve the system without undue risk to participating carriers. We feel that this is an achievable goal that requires awareness by the regulatory agency of the need for such extensive evaluation data at the individual level, sensitivity to the fact that such information represents a potentially ticking time bomb while it resides in files that may have to be released to the public, and a willingness to shield such information from outside attack. NASA has achieved this protection in the Aviation Safety Reporting System; it should be possible, under a cost-benefit analysis, to develop procedures and safeguards that are equitable to all and will allow the most effective utilization of data.

References

FAA. Federal Aviation Administration Release 83-85, 1985.

Foushee, H.C. Dyads and triads at 35,000 feet: Factors affecting group process and aircrew performance. American Psychologist, 1984, 39, 885-993.

Foushee, H.C. & Manos, K.L. Information trasfer within the cockpit: Problems in intracockpit communications. In C.E. Billings & E.S. Cheaney (Eds.) Information transfer problems in the aviation system. NASA Report No. TP-1875. Moffett Field, CA: NASA-Ames Research Center, 1981.

Gerathewohl, S.J. Age and the aviator: Developing a psychophysiological proficiency index for pilots. Wehrpsychologische Untersuchungen, 1978, 13, 15-26.

Goeters, K.M. Selection of personnel for complex operations as demonstrated by pilot selection. Paper presenter to the Ispra-Course "Training of operating personnel for technologies with hazard potential: Theory and practice." Ispra (Italy), 1980.

Hackman, J.R. A normative model of work team effectiveness. Technical report No. 2, Research program on group effectiveness, New Haven: Yale School of Organization and Management, 1983.

Harper, C.R., Kidera, G.J., & Cullen, J.F. Study of simulated airline pilot incapacitation: Phase II, Subtle or partial loss of function. Aerospace Medicine, 1971, 42.

Helmreich, R.L. Applied social psychology: The unfulfilled promise. Personality and Social Psychology Bulletin, 1975, 1, 548-561.

Helmreich, R.L. Pilot selection and training, Paper presented at the American Psychological Association, annual meeting, Washington, D.C., 1982.

Helmreich, R. L. What changes and what endures: The capabilities and limitations of training and selection. In N. Johnston (Ed.) Proceedings of the Aer Lingus/Irish Airline Pilots Association Flight Symposium, Dublin, Ireland, 1983.

Helmreich, R.L., Bakeman, R.A., & Scherwitz, L., The study of small groups. In Musson, P. (Ed.), Annual Review of Psychology. Palo Alto: Review Press, 1973, 337-351.

Helmreich, R.L. & Spence, J.T. The Work and Family Orientation Questionnaire: An objective instrument to assess components of achievement motivation and attitudes toward family and career. JSAS Catalog of selected documents in psychology, 8, 35, Ms 1677.

Kirsch, H., Goeters, K.M., & Ewe, R. Faktoranalyse eines neuen mehrdimensionalen Personlichkeits-Fragebogens. DFVLR FB, 1975, 75-120.

Lauber, J.K. & Foushee, H.C. Guidelines for line oriented flight training Vol 1. NASA Report No. CP-2184. Moffett Field, CA: NASA-Ames Research Center, 1981.

Mearns, D.J. FDR - the pilot's friend: BA/BALPA co-operation in action. In N. Johnston (Ed.) Proceedings of the Aer Lingus/Irish Airline Pilots Association Flight Symposium, Dublin, Ireland, 1983.

Melton, A.W. Army air forces. Aviation psychology program. Report No. 4. Apparatus Tests. Washington. Defense Documentation Center, 1947.

National Transportation Safety Board. Aircraft accident report (NTSB Report No. AAR-82-8) Washington, D.C.: NTSB Bureau of Accident Investigation, 1982.

National Transportation Safety Board. Aircraft accident report (NTSB Report No. AAR-80-1). Washington, D.C.: NTSB Bureau of Accident Investigation, 1980.

Ruffell Smith, H.P. A simulator study of the interaction of pilot workload with errors, vigilance, and decisions. NASA Report No. TM-78482. Moffett Field, CA: NASA-Ames Research Center, 1979.

Wittl, H. Schatzungen von Testvalidataten an stichproben mit eingeschrankter Varianz am Beispiel von Flugaeugfuhrereignungstests. Z. exp. angew. Psychologie, 1970, 17, 158-181.